

STATISZTIKA DATA SCIENCE-HEZ

KORRELÁCIÓ ELEMZÉS ALAPISMERETEK

Korrelációs együtthatók számítása, alkalmazása

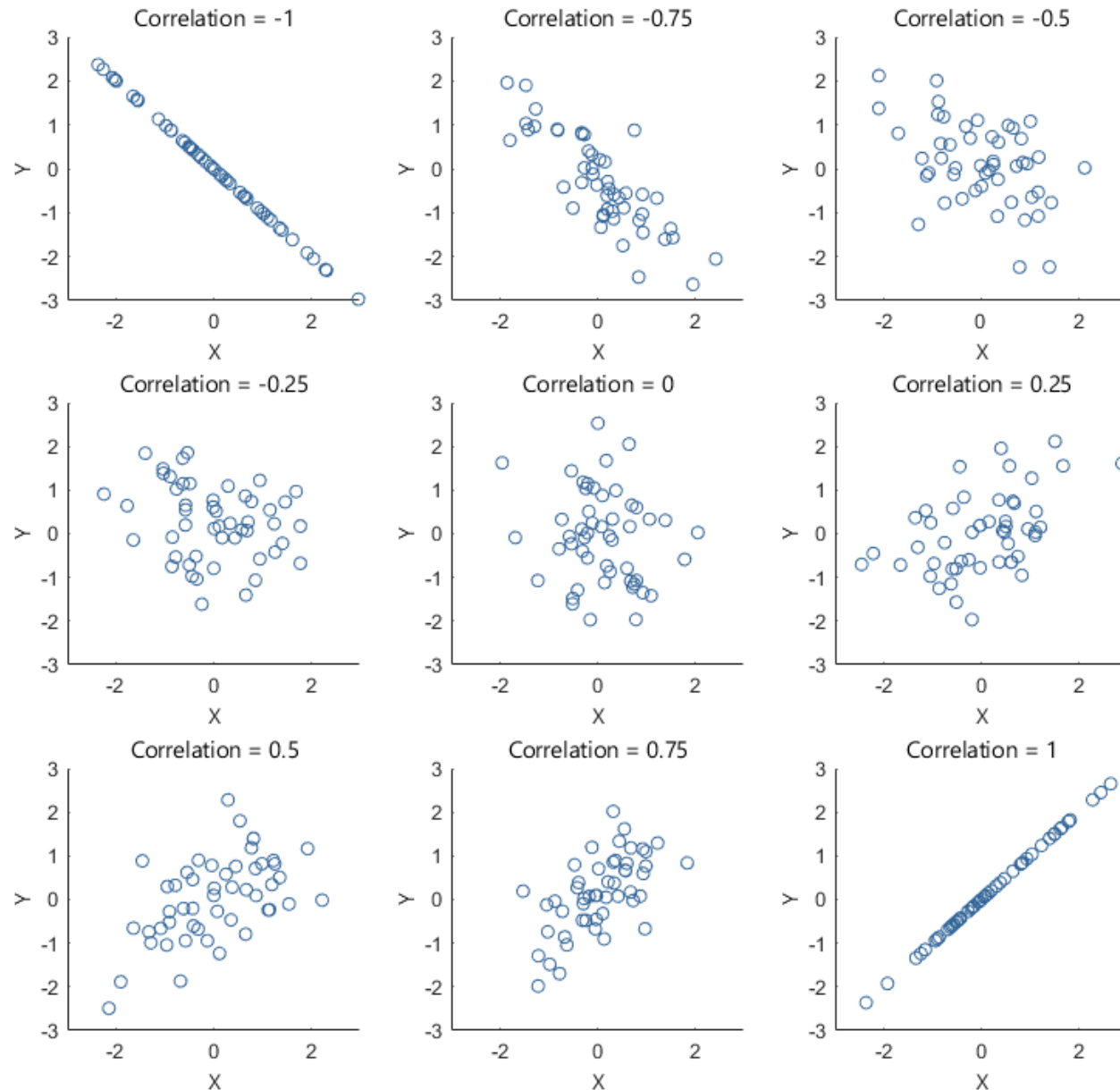
Korreláció és szignifikancia

Korreláció vizualizációja

Korreláció VS ok-okozati összefüggés



Realizations of couples of random variables X and Y
with different correlation coefficients



MI AZ A KORRELÁCIÓ ELEMZÉS?

Egy statisztikai módszer,
amely két változó
közti kapcsolat

erősségének és
irányának

meghatározására
használható.



PEARSON-FÉLE KORRELÁCIÓS EGYÜTTHATÓ

- A Pearson korrelációs együttható (r) a lineáris kapcsolat erősségét és irányát méri két folytonos és normál eloszlású numerikus változó között.
- Értéke -1 és +1 között mozog, ahol -1 tökéletes negatív lineáris kapcsolatot, +1 pedig tökéletes pozitív lineáris kapcsolatot jelez.
- Az $r = 0$ azt jelzi, hogy nincs lineáris kapcsolat.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

\bar{x} = mean of values in x variable

\bar{y} = mean of values in y variable



PEARSON-FÉLE KORRELÁCIÓS EGYÜTTHATÓ

- A Pearson korrelációs együttható (r) a lineáris kapcsolat erősségét és irányát méri két folytonos és normál eloszlású numerikus változó között.

lineáris
kapcsolat

folytonos
változók

normál
eloszlású
változók

numerikus
változók

JDSA_2024_correlation_analyses.ipynb



PEARSON-FÉLE KORRELÁCIÓ LIMITÁCIÓJA

**lineáris
kapcsolat**

*Mi van, ha nem
lineáris?
(hatvány, exp, log)*

**folytonos
változók**

*Mi van, ha
diszkrét
változóink (is)
vannak?*

**normál
eloszlású
változók**

*Mi van, ha
vannak kiugró
értékek?
Mi van ha nem
az átlag körül
csúcsodik a
görbe?*

**numerikus
változók**

*Mi van, ha
kategórikus
változóink (is)
vannak?*



KORRELÁCIÓ SZÁMÍTÁS MÁS MÓDJAI

- ***Spearman-féle rangkorrelációs együttható***
- ***Kendall's Tau rangkorrelációs együttható***
- ***Chi-négyzet teszt***
- ***Fisher pontos teszt***
- ***Kappa együttható***
- ***Pontbiseriális korreláció***
- ***ANOVA teszt***
- ***Cramér V statisztika***
- ***Kruskal-Wallis teszt***



SPEARMAN-FÉLE RANGKORRELÁCIÓS EGYÜTTHATÓ

- A Spearman-féle rangkorrelációs együttható (r_s vagy ρ) egy nem paraméteres teszt, amely két változó közötti monoton kapcsolat erősségét méri.
- A változók rangjain alapul
- Akkor használjuk, amikor a változók nem feltétlenül mutatnak lineáris kapcsolatot, vagy az adatok között vannak ordinálisak (sorba rendezhető kategórikus adatok).
- Az értéke szintén -1 és +1 között mozog.

monoton
kapcsolat

változók lehetnek
numerikusak vagy
ordinálisak

változók
rangjain alapul

JDSA correlation analyses.ipynb



SPEARMAN-FÉLE RANGKORRELÁCIÓS EGYÜTTHATÓ

***Spearman-féle rangkorrelációs
együttható (r_s vagy ρ)***

=

***Pearson-féle korrelációs
együttható (r)***

***nyersadatok helyett
a rangokon alkalmazva***

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

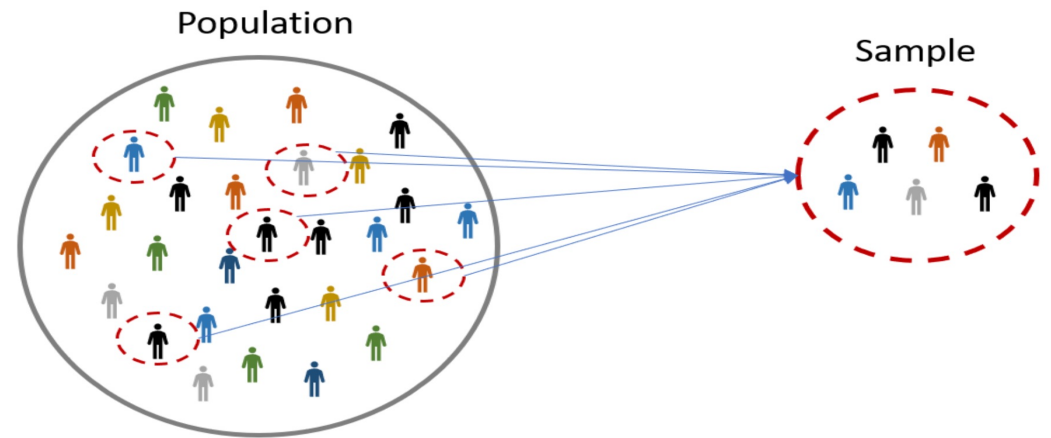
n = number of observations



SZIGNIFIKANCIA KÉRDÉSE

*Mennyire a véletlen műve a
kapott eredmény?*

*Mennyire általánosítható a
kapott eredmény?*





SZIGNIFIKANCIA KÉRDÉSE

JDSA_2024_correlation_analyses.ipynb

1) **Null hipotézis (H_0) = nincs lineáris kapcsolat**

2) **korrelációs együttható (r) meghatározása
+ következtetések levonása**

3) **p -érték meghatározása**

= annak a valószínűsége, hogy a kapott r érték
csupán a véletlen műve (és mégis a H_0 az igaz)

t-teszt és szabadsági fokok
meghatározása után
egy eloszlás táblázatból

4) **ha $p < 0,05^*$**

= elutasítjuk a H_0 -t \Rightarrow van lineáris kapcsolat

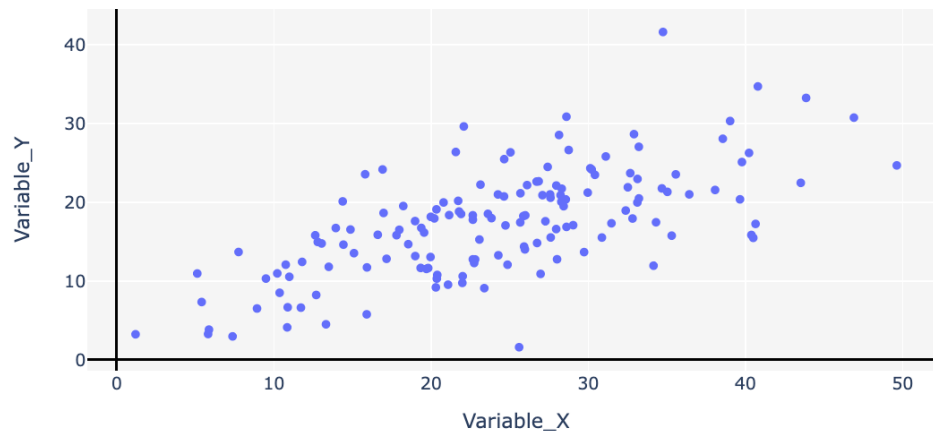
ha $p > 0,05$

= (most) nem tudjuk elutasítani a H_0 -t
 \Rightarrow nincs lineáris kapcsolat

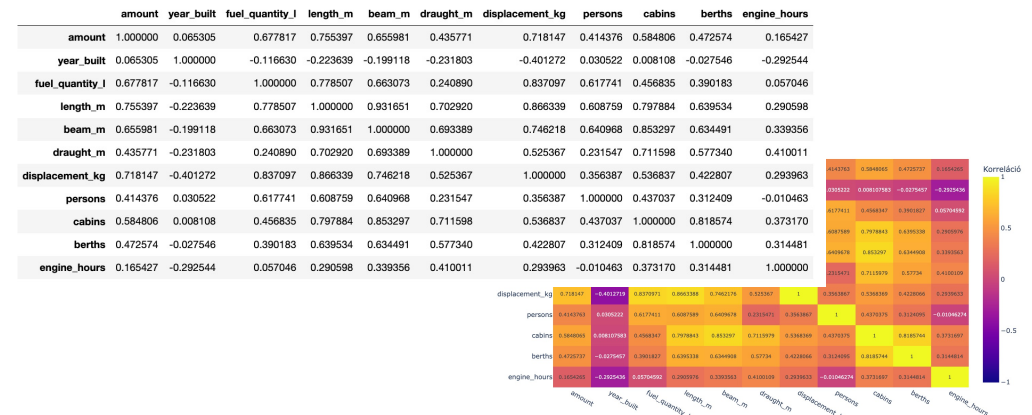
**: 5% kockázatot vállalunk rá, hogy úgy vetjük el a H_0 -t, hogy az mégis igaz*

KORRELÁCIÓ VIZUALIZÁCIÓJA

- *scatter plotok*
= *szórás diagramok*



- *korrelációs mátrix*
heatmap-ekkel kiegészítve



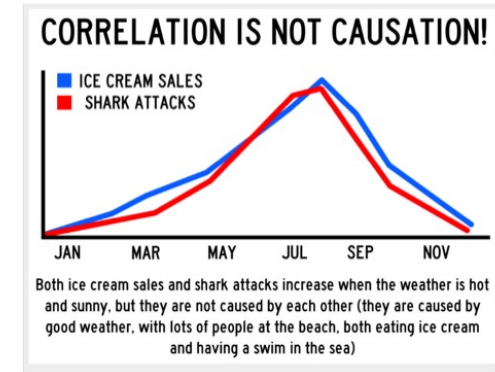
JDSA_2024_correlation_analyses.ipynb





Korreláció elemzés

= van/nincs kapcsolat a változók között



DE már ez is fontos és elegendés lehet:

- további elemzésekhez

pl. gépi tanulós modellek bemeneti változóinak meghatározásához

- hipotézis alkotáshoz

aminek ok-okozati viszonyát aztán további kutatások fel tudják tárn

THANKS FOR LISTENING



ANY QUESTIONS?



AJÁNLOTT / FELHASZNÁLT IRODALOM

Ajánlott irodalom

O'Reilly kiadó *Practical Statistics for Data Scientists*

David Borman: *Statistic 101*

***Correlation \neq Causation* - további érdekes példák**

Képek forrása

<https://www.statlect.com/fundamentals-of-probability/linear-correlation>

<https://medium.com/@abdallahshraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9>

https://www.researchgate.net/figure/Correlation-Coefficient-and-Strength-of_tbl1_339336406

<https://datascientistinterviews.quora.com/Difference-between-Sample-and-Population>

<https://anyi-guo.medium.com/correlation-pearson-vs-spearman-cl5e581cl2ce>

<https://stock.adobe.com/images/causation-correlation/240972211>

